

COUNTING AND LOCATING THE SOLUTIONS OF POLYNOMIAL SYSTEMS OF MAXIMUM LIKELIHOOD EQUATIONS, II: THE BEHRENS-FISHER PROBLEM

Max-Louis G. Buot, Serkan Hoşten, and Donald St. P. Richards

Xavier University, San Francisco State University,

Penn State University and SAMSI

Abstract: Let μ be a p -dimensional vector, and let Σ_1 and Σ_2 be $p \times p$ positive definite covariance matrices. On being given random samples of sizes N_1 and N_2 from independent multivariate normal populations $N_p(\mu, \Sigma_1)$ and $N_p(\mu, \Sigma_2)$, respectively, the Behrens-Fisher problem is to solve the likelihood equations for estimating the unknown parameters μ , Σ_1 , and Σ_2 . We shall prove that for $N_1, N_2 > p$ there are, almost surely, exactly $2p + 1$ complex solutions of the likelihood equations. For the case in which $p = 2$, we utilize Monte Carlo simulation to estimate the relative frequency with which a typical Behrens-Fisher problem has multiple real solutions; we find that multiple real solutions occur infrequently.

Key words and phrases: Behrens-Fisher problem, Bézout's theorem, maximum likelihood estimation, maximum likelihood degree.

1 Introduction

Let $\mu \in \mathbb{R}^p$ be a p -dimensional vector, and let Σ_1 and Σ_2 be $p \times p$ positive definite (symmetric) matrices. Consider independent multivariate normal populations, $N_p(\mu, \Sigma_1)$ and $N_p(\mu, \Sigma_2)$, from which we have been given random samples X_1, \dots, X_{N_1} and Y_1, \dots, Y_{N_2} , respectively. On the basis of the given data, the famous Behrens-Fisher problem (Behrens (1929), Fisher (1939)) is to estimate the parameters μ , Σ_1 , and Σ_2 by means of the method of maximum likelihood.

It is well-known that the corresponding system of likelihood equations cannot be solved explicitly, and that has led many to propose alternative solutions to the Behrens-Fisher problem (Anderson, 2003, p. 187 ff.). More importantly, the Behrens-Fisher problem is an early example of a hypothesis testing problem

involving exponential families of densities and for which the resulting sufficient statistics, when the parameters are restricted to the parameter space determined by H_0 , fail to be complete (Linnik, 1967). In such a situation, nuisance parameters exist, and the construction of an exact size- α test is a difficult problem.

Consequently, the literature on the Behrens-Fisher problem is substantial, reflecting the intense interest which the problem has generated since its inception. Indeed, the problem has generated an extensive philosophical discussion as well as many efforts to derive solutions which are optimal for statistical inference (Wallace (1980), Kim and Cohen (1998), Stuart and Ord (1994)). In this paper, we determine the number of solutions of the likelihood equations.

For the case in which $p = 1$, there are three unknown scalar parameters, *viz.*, μ , the common mean, and σ_1^2 and σ_2^2 , the population variances. In this case, Sugiura and Gupta (1987) reduced the system of equations to a cubic equation in μ and deduced that, almost surely, there are three complex solutions; they observed also that the likelihood equation tended to have multiple real solutions if σ_1^2 and σ_2^2 are small in comparison with μ , and otherwise that the likelihood equation usually has a unique real solution. Drton (2007) also studied the univariate Behrens-Fisher problem and showed, in particular, that if the null hypothesis is true then the probability of multiple real solutions tends to zero as the sample sizes tend to infinity. We will prove the analogous result for the multivariate Behrens-Fisher problem in Theorem 4.1.

In this paper, as in the article of Buot and Richards (2006), we apply results from the theory of algebraic geometry to study the solution set of the system of likelihood equations for the multivariate Behrens-Fisher problem. Generalizing the univariate result described earlier, we shall prove the following result:

Theorem 1.1. *Suppose that $N_1, N_2 > p$. Then, almost surely, there are exactly $2p + 1$ complex solutions of the system of likelihood equations for the multivariate Behrens-Fisher problem. In particular, almost surely, there always exists at least one real solution.*

2 Derivation of the likelihood equations

Denote by \bar{X} and \bar{Y} the means of the samples from $N_p(\mu, \Sigma_1)$ and $N_p(\mu, \Sigma_2)$, respectively. By standard calculations (cf., Mardia, *et al.* (1979), p. 142), we find that the likelihood equations for estimating μ , Σ_1 and Σ_2 are:

$$\begin{aligned}\hat{\Sigma}_1 &= N_1^{-1} \sum_{j=1}^{N_1} (X_j - \hat{\mu})(X_j - \hat{\mu})', \\ \hat{\Sigma}_2 &= N_2^{-1} \sum_{j=1}^{N_2} (Y_j - \hat{\mu})(Y_j - \hat{\mu})'\end{aligned}\tag{2.1}$$

and

$$(N_1 \hat{\Sigma}_1^{-1} + N_2 \hat{\Sigma}_2^{-1}) \hat{\mu} = N_1 \hat{\Sigma}_1^{-1} \bar{X} + N_2 \hat{\Sigma}_2^{-1} \bar{Y}.\tag{2.2}$$

Some authors have proposed the following iterative algorithm for solving (2.1) and (2.2):

- (1) Begin the iteration with initial estimates $\hat{\Sigma}_{i,0} = \tilde{S}_i$, $i = 1, 2$, where

$$\begin{aligned}\tilde{S}_1 &= N_1^{-1} \sum_{j=1}^{N_1} (X_j - \bar{X})(X_j - \bar{X})', \\ \tilde{S}_2 &= N_2^{-1} \sum_{j=1}^{N_2} (Y_j - \bar{Y})(Y_j - \bar{Y})'. \end{aligned}\tag{2.3}$$

- (2) Apply (2.2) to calculate $\hat{\mu}_0$, the corresponding estimate of μ , in the form

$$\hat{\mu}_0 = (N_1 \hat{\Sigma}_{1,0}^{-1} + N_2 \hat{\Sigma}_{2,0}^{-1})^{-1} (N_1 \hat{\Sigma}_{1,0}^{-1} \bar{X} + N_2 \hat{\Sigma}_{2,0}^{-1} \bar{Y}).$$

- (3) Use the value of $\hat{\mu}_0$ obtained in Step (2) to calculate $\hat{\Sigma}_{i,1}$, an updated value of $\hat{\Sigma}_{i,0}$, using the formulas

$$\hat{\Sigma}_{1,1} = \tilde{S}_1 + (\bar{X} - \hat{\mu}_0)(\bar{X} - \hat{\mu}_0)', \quad \hat{\Sigma}_{2,1} = \tilde{S}_2 + (\bar{Y} - \hat{\mu}_0)(\bar{Y} - \hat{\mu}_0)',$$

which are a consequence of (2.5) and (2.6) below.

- (4) Return to Step (2) and update $\hat{\mu}_j$ until the sequences $\hat{\Sigma}_{1,j}$ and $\hat{\Sigma}_{2,j}$, $j = 1, 2, 3, \dots$, converge.

We are grateful to Mathias Drton for pointing out that his work in Drton and Eichler (2006) implies that this algorithm converges to a saddle point or a local (but not necessarily a global) maximum of the likelihood function. If the likelihood function were found to be multimodal, a phenomenon which has been encountered recently by Drton and Richardson (2004) in a study of seemingly unrelated regression models, then any numerical algorithm for solving the system of likelihood equations necessarily must include some information about the choice of initial values.

At first glance, the likelihood equations may appear to be a system of $p(p+2)$ equations in $p(p+2)$ variables comprising the p components of μ and the $p(p+1)/2$ entries of both Σ_1 and Σ_2 . However, a closer inspection of (2.1) and (2.2) reveals that if $\hat{\mu}$ is known then $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are determined completely. We shall show later how to eliminate $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ from (2.2) to obtain a system of p cubic equations in the variables $\hat{\mu}_1, \dots, \hat{\mu}_p$.

Proposition 2.1. *The likelihood equations (2.1) and (2.2) for the Behrens-Fisher problem are equivalent to*

$$\frac{N_1 \tilde{S}_1^{-1}(\bar{X} - \hat{\mu})}{1 + (\bar{X} - \hat{\mu})' \tilde{S}_1^{-1}(\bar{X} - \hat{\mu})} + \frac{N_2 \tilde{S}_2^{-1}(\bar{Y} - \hat{\mu})}{1 + (\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1}(\bar{Y} - \hat{\mu})} = 0. \quad (2.4)$$

Proof. We apply to the sums in (2.1) the standard procedure of writing each term $X_i - \hat{\mu}$ as $X_i - \bar{X} + \bar{X} - \hat{\mu}$, and similarly for each term $Y_i - \hat{\mu}$. This leads to the formulas

$$\hat{\Sigma}_1 = \tilde{S}_1 + (\bar{X} - \hat{\mu})(\bar{X} - \hat{\mu})' \quad (2.5)$$

and

$$\hat{\Sigma}_2 = \tilde{S}_2 + (\bar{Y} - \hat{\mu})(\bar{Y} - \hat{\mu})' \quad (2.6)$$

where \tilde{S}_1 and \tilde{S}_2 are defined in (2.3). By a special case of Woodbury's theorem (cf., Muirhead (1982), p. 580, Theorem A5.1) we have, for any nonsingular $p \times p$ matrix M and any column vector $v \in \mathbb{R}^p$,

$$(M + vv')^{-1} = M^{-1} - \frac{M^{-1}vv'M^{-1}}{1 + v'M^{-1}v}.$$

Multiplying the latter equation on each side from the right by v and simplifying

the result, we obtain

$$\begin{aligned}
(M + vv')^{-1}v &= M^{-1}v - \frac{M^{-1}vv'M^{-1}v}{1 + v'M^{-1}v} \\
&= \frac{(1 + v'M^{-1}v)M^{-1}v - (M^{-1}v)(v'M^{-1}v)}{1 + v'M^{-1}v} \\
&= \frac{M^{-1}v}{1 + v'M^{-1}v}.
\end{aligned}$$

Setting $M = \tilde{S}_1$ and $v = \bar{X} - \hat{\mu}$, we obtain

$$\begin{aligned}
\hat{\Sigma}_1^{-1}(\bar{X} - \hat{\mu}) &\equiv (\tilde{S}_1 + (\bar{X} - \hat{\mu})(\bar{X} - \hat{\mu})')^{-1}(\bar{X} - \hat{\mu}) \\
&= \frac{\tilde{S}_1^{-1}(\bar{X} - \hat{\mu})}{1 + (\bar{X} - \hat{\mu})'\tilde{S}_1^{-1}(\bar{X} - \hat{\mu})},
\end{aligned} \tag{2.7}$$

and, similarly,

$$\hat{\Sigma}_2^{-1}(\bar{Y} - \hat{\mu}) = \frac{\tilde{S}_2^{-1}(\bar{Y} - \hat{\mu})}{1 + (\bar{Y} - \hat{\mu})'\tilde{S}_2^{-1}(\bar{Y} - \hat{\mu})}. \tag{2.8}$$

On rewriting (2.2) as

$$N_1\hat{\Sigma}_1^{-1}(\bar{X} - \hat{\mu}) + N_2\hat{\Sigma}_2^{-1}(\bar{Y} - \hat{\mu}) = 0,$$

it follows from (2.7) and (2.8) that (2.2) is equivalent to (2.4). \square

3 The maximum likelihood degree of the Behrens-Fisher problem

Following Catanese, *et al.* (2006) and Hosten, *et al.* (2005) we will call the number of complex solutions to the likelihood equations the *maximum likelihood degree*. In this section we shall prove Theorem 1.1, namely, that the maximum likelihood (or ML) degree of the Behrens-Fisher problem is $2p + 1$. Before providing the details of the proof, it is instructive to understand why the theorem holds for $p = 1$ and $p = 2$. Let us denote by $D_X(\hat{\mu})$ and $D_Y(\hat{\mu})$ the denominators $1 + (\bar{X} - \hat{\mu})'\tilde{S}_1^{-1}(\bar{X} - \hat{\mu})$ and $1 + (\bar{Y} - \hat{\mu})'\tilde{S}_2^{-1}(\bar{Y} - \hat{\mu})$, respectively, which appear in the likelihood equations (2.4).

Lemma 3.1. *Let*

$$N_1D_Y(\hat{\mu})\tilde{S}_1^{-1}(\bar{X} - \hat{\mu}) + N_2D_X(\hat{\mu})\tilde{S}_2^{-1}(\bar{Y} - \hat{\mu}) = 0 \tag{3.1}$$

be the system of polynomial equations obtained by clearing denominators in the likelihood equations (2.4), and suppose that $\hat{\mu}$ is a solution to (3.1). Then $D_X(\hat{\mu}) = 0$ if and only if $D_Y(\hat{\mu}) = 0$.

Proof. Suppose that $D_X(\hat{\mu}) = 0$. On multiplying (3.1) from the left by $(\bar{X} - \hat{\mu})'$ we obtain $N_1 D_Y(\hat{\mu})(D_X(\hat{\mu}) - 1) = 0$ and so we deduce that $D_Y = 0$. Similarly, starting with the assumption that $D_Y = 0$, we deduce that $D_X = 0$. \square

We remark that, because $D_X(\hat{\mu})$ and $D_Y(\hat{\mu})$ are strictly positive for any real $\hat{\mu}$, the system of equations (2.4) and (3.1) are equivalent when determining *real* solutions only. However, in the calculation of *complex* solutions, the likelihood equations (2.4) are not equivalent to (3.1) since it is possible that the denominators are zero for complex $\hat{\mu}$.

Let J be the ideal defined by the equations (3.1) and let $I = \langle D_X(\hat{\mu}), D_Y(\hat{\mu}) \rangle$ be the ideal of zeros common to the first and second denominators. Then we need to compute and count the solutions to $J : I$. For the case in which $p = 1$ there is a single univariate cubic polynomial in (3.1) which, generically, has three complex roots. Since two generic univariate polynomials (in this case, $D_X(\hat{\mu})$ and $D_Y(\hat{\mu})$) have no common roots then the ideal I has, in general, no solutions. Hence we conclude that $J : I$ has exactly three solutions for the case in which $p = 1$.

We now consider the case in which $p = 2$. Since two quadrics in two variables have, generically, four complex roots then there are four generic solutions to I . Similarly, since two cubics in two variables have generically nine complex roots then there are nine generic solutions to J . Therefore $J : I$ has five complex roots for the case in which $p = 2$.

Unfortunately, this counting argument fails even for $p = 3$. In this case, we have two quadrics in three variables, so there are infinitely many solutions to I and hence also to J . Yet, $J : I$ still has finitely many solutions. Theorem 1.1 follows from the following result of Catanese, *et al.* (2006).

Theorem 3.2. (Catanese, *et al.*, 2006) *Let f_1, \dots, f_n be polynomials of degrees b_1, \dots, b_n , respectively, in the variables x_1, \dots, x_d ; let u_1, \dots, u_n be integers; let $f = f_1^{u_1} \dots f_n^{u_n}$; and consider the critical equations*

$$\frac{1}{f} \frac{\partial f}{\partial x_1} = \frac{1}{f} \frac{\partial f}{\partial x_2} = \dots = \frac{1}{f} \frac{\partial f}{\partial x_n} = 0$$

of $\log f = \sum_{i=1}^n u_i \log f_i$. If the number of complex solutions to this system of equations is finite then that number is less than or equal to the coefficient of z^d in the generating function

$$\frac{(1-z)^d}{(1-b_1z)(1-b_2z)\cdots(1-b_nz)}.$$

Equality holds if the coefficients of the polynomials f_i are sufficiently generic.

Before we proceed, there are a few points that need clarification in Theorem 3.2. First of all, given the integers b_1, \dots, b_n there exists a fixed polynomial $G = G_{b_1, \dots, b_n}$ in the coefficients of n polynomials in d variables with degrees b_1, \dots, b_n so that we call f_1, \dots, f_n *generic* if $G(f_1, \dots, f_n) \neq 0$. Furthermore, when f_1, \dots, f_n is generic in this sense the number of complex solutions to the critical equations is given by the formula in the statement of the theorem. In other words, genericity already implies the finiteness of the number of complex solutions. This follows from Theorem 5 in Catanese, *et al.* (2006) leading up to the proof of Theorem 3.2.

Before providing the proof of Theorem 1.1, we show that the coefficients of $D_X(\hat{\mu})$ and $D_Y(\hat{\mu})$ are generic for almost all data X_1, \dots, X_{N_1} and Y_1, \dots, Y_{N_2} . First we need the following result which has a standard proof in the literature (for instance based on the argument on page 76 in Anderson (2003)). We present our own proof.

Lemma 3.3. *Suppose that $N + 1 > p$. Then given any $p \times p$ positive definite matrix S there exist $X_1, \dots, X_{N+1} \in \mathbb{R}^p$ such that $S = \sum_{i=1}^{N+1} (X_i - \bar{X})(X_i - \bar{X})'$.*

Proof. We can assume, without loss of generality that $\bar{X} = 0$. Now let $X_i = (X_{i1}, \dots, X_{ip})'$ for $i = 1, \dots, N$ and let $X_{N+1} = -\sum_{i=1}^N X_i$. Given a positive definite matrix S , there exists a nonsingular symmetric matrix U such that $USU' = \Lambda$ where Λ is a diagonal matrix with diagonal entries $\lambda_i > 0$, $i = 1, \dots, p$. Hence it is enough to prove the result for diagonal matrices Λ . The required identity $\Lambda = \sum_{i=1}^{N+1} X_i X_i'$ gives rise to $p(p+1)/2$ polynomial equations, namely,

$$\sum_{i=1}^N \sum_{j=1}^N X_{ik} X_{jk} = \frac{\lambda_k}{2},$$

$k = 1, \dots, p$, and

$$X_{1i}(X_{1j} + \sum_{k=1}^N X_{kj}) + X_{2i}(X_{2j} + \sum_{k=1}^N X_{kj}) + \dots + X_{Ni}(X_{Nj} + \sum_{k=1}^N X_{kj}) = 0,$$

$1 \leq i < j \leq p$. We claim that there exists at least one real solution to the above system where $X_{ij} = 0$ for $j = 1, \dots, p$ and $i = j + 1, \dots, N$. It is easy to check that $X_{11} = \sqrt{\frac{\lambda_1}{2}}$ with $X_{1k} = X_{2k} = \dots = X_{k-1,k} = \sqrt{\frac{\lambda_k}{k(k+1)}}$ and $X_{kk} = -k\sqrt{\frac{\lambda_k}{k(k+1)}}$ for $k = 2, \dots, p$ give such a solution. \square

Theorem 3.4. *For generic data X_1, \dots, X_{N_1} and Y_1, \dots, Y_{N_2} the denominators D_X and D_Y are generic.*

Proof. D_X and D_Y are quadratic forms in p variables. In the light of our remarks after Theorem 3.2 there exists a fixed polynomial G in the coefficients of two quadratic forms in p variables such that D_X and D_Y are generic if $G(D_X, D_Y) \neq 0$. We need to show that this condition holds for generic data. As in the proof of Lemma 3.3, the entries of \tilde{S}_1 and \tilde{S}_2 are polynomials in the data. The same lemma implies that the polynomial maps defined from the data spaces $\mathbb{R}^{p \times N_1}$ and $\mathbb{R}^{p \times N_2}$ are surjective onto the cone of semidefinite matrices in $\mathbb{R}^{p(p+1)/2}$. Therefore there exist data vectors X_1, \dots, X_{N_1} and Y_1, \dots, Y_{N_2} such that $G(D_X, D_Y) \neq 0$. If the statement in the theorem is not true, then there exists a Zariski open subset $U \subset \mathbb{R}^{p \times N_1} \times \mathbb{R}^{p \times N_2}$ such that for all $(X_1, \dots, X_{N_1} : Y_1, \dots, Y_{N_2}) \in U$ we have $G(D_X, D_Y) = 0$. But this means that G is identically zero, and this is a contradiction. \square

Proof of Theorem 1.1: Denoting by $L(\mu, \Sigma_1, \Sigma_2)$ the likelihood function for the Behrens-Fisher problem, then it is well-known that

$$L(\hat{\mu}, \hat{\Sigma}_1, \hat{\Sigma}_2) = (2\pi e)^{-(N_1+N_2)p/2} |\hat{\Sigma}_1|^{-N_1/2} |\hat{\Sigma}_2|^{-N_2/2}.$$

By (2.5) and (2.6), we have $|\hat{\Sigma}_1| = |\tilde{S}_1| \cdot D_X(\hat{\mu})$ and $|\hat{\Sigma}_2| = |\tilde{S}_2| \cdot D_Y(\hat{\mu})$. Therefore

$$\begin{aligned} L(\hat{\mu}, \hat{\Sigma}_1, \hat{\Sigma}_2) \\ = (2\pi e)^{-(N_1+N_2)p/2} |\tilde{S}_1|^{-N_1/2} |\tilde{S}_2|^{-N_2/2} (D_X(\hat{\mu}))^{-N_1/2} (D_Y(\hat{\mu}))^{-N_2/2}. \end{aligned}$$

It now is clear that, to find the maximum value of L , we need to minimize

$$(1 + (\bar{X} - \hat{\mu})' \tilde{S}_1^{-1} (\bar{X} - \hat{\mu}))^{N_1/2} (1 + (\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu}))^{N_2/2}. \quad (3.2)$$

Equivalently, we may minimize the logarithm of this expression, and since the critical equations of the logarithm of (3.2) are precisely the likelihood equations in Proposition 2.1, then Theorem 3.2 implies that the maximum likelihood degree of the Behrens-Fisher problem is equal to the coefficient of z^p in the power series expansion of the rational function

$$\frac{(1-z)^p}{(1-2z)^2},$$

provided that the data X_1, \dots, X_{N_1} and Y_1, \dots, Y_{N_2} , and hence D_X and D_Y are generic. By expanding this rational function in a power series in z , we find that this coefficient equals

$$\sum_{i+j=p} (-1)^i 2^j \binom{p}{i} (j+1),$$

and an elementary calculation shows that this sum equals $2p+1$.

Step 4 in Algorithm 7 in Hoşten, *et al.* (2005) and the theory of Gröbner bases imply that all $2p+1$ complex solutions can be obtained from the roots of a univariate polynomial of degree $2p+1$. Since $D_X(\hat{\mu})$ and $D_Y(\hat{\mu})$ have real coefficients, this univariate polynomial also has real coefficients. In particular, since roots occur in complex conjugate pairs then at least one root is real. \square

Remark 3.5. We note that our arguments which led to the derivation of the ML degree of the Behrens-Fisher problem also apply to the more general problem of multivariate analysis of variance (MANOVA). Suppose that we have independent multivariate normal populations $N_p(\mu, \Sigma_1), \dots, N_p(\mu, \Sigma_{k+1})$ and that, on the basis of random samples from each population, we wish to derive the maximum likelihood estimators of the parameters μ and $\Sigma_1, \dots, \Sigma_{k+1}$. By arguments similar to those in Section 2, we obtain analogous likelihood equations as in Proposition 2.1 where now there are $k+1$ rational summands in each of the p equations. It then follows from Theorem 3.2 that the ML degree for the MANOVA problem is

$$d(k, p) := \sum_{i+j=p} (-1)^i 2^j \binom{p}{i} \binom{j+k}{k}. \quad (3.3)$$

By writing this result in the form

$$d(k, p) = 1 + \sum_{j=1}^p (-1)^{p-j} 2^j \binom{p}{j} \binom{j+k}{k},$$

we find that $d(k, p)$ is odd; therefore, there always exists a real solution to the system of likelihood equations.

We note that $d(k, p)$ can be evaluated using methods from the calculation of combinatorial sums, as follows: First, we write

$$2^j \binom{j+k}{k} = \frac{1}{k!} \left(\frac{d}{dt} \right)^k t^{j+k} \Big|_{t=2}.$$

Inserting this formula in the sum in (3.3) and interchanging derivatives and summation, we obtain

$$\begin{aligned} d(k, p) &= \frac{1}{k!} \left(\frac{d}{dt} \right)^k t^k \sum_{j=0}^p (-1)^{p-j} \binom{p}{j} t^j \Big|_{t=2} \\ &= \frac{1}{k!} \left(\frac{d}{dt} \right)^k t^k (t-1)^p \Big|_{t=2}. \end{aligned} \quad (3.4)$$

In particular, $d(1, p) = 2p+1$, the ML degree of the Behrens-Fisher problem, and $d(2, p) = 2p(p+1)+1$. The general formula for $d(k, p)$ is interesting even in the case $p=1$, for it yields the ML degree of the one-dimensional $(k+1)$ -population MANOVA problem to be $2k+1$. Further, by substituting $t = (1+u)/2$ in (3.4), we recognize the outcome as Rodrigues' formula (Szegő, 1939, p. 66) for a Jacobi polynomial $P_k^{(p-k,0)}$, and we obtain $d(k, p) = P_k^{(p-k,0)}(3)$, $p \geq k$.

4 Simulations and a large sample size result

Having determined the number of solutions of the system of likelihood equations (3.1) it is natural to seek the number of *real* solutions, for it is those solutions which are of interest in statistical inference. Not surprisingly, it appears to be difficult to determine an algebraic expression for the number of real solutions of the system; indeed, this is also the case for the general theory of systems of polynomial equations.

To study the real solutions of the system (3.1), we considered the case in which $p=2$, presenting empirical evidence that multiple solutions occur rarely if the model is correctly specified. In each simulation run, we first used a random number generator to generate sample sizes N_1 and N_2 , and a mean vector μ . We next generated lower triangular matrices T_1 and T_2 with positive diagonal

entries, after which we set $\Sigma_k = T_k T_k'$, $k = 1, 2$. Finally, we simulated a random sample of vectors Z_1, \dots, Z_{N_1} from $N_2(0, I_2)$, and then we set $X_j = T_1 Z_j + \mu$, $j = 1, \dots, N_1$. It follows from standard distribution theory that X_1, \dots, X_{N_1} constitutes a simulated sample from the bivariate normal population $N_2(\mu, \Sigma_1)$. In a similar manner, we simulated an independent random sample Y_1, \dots, Y_{N_1} from $N_2(\mu, \Sigma_2)$.

The solutions of the resulting likelihood equations (3.1) were computed numerically using **PHCpack** (Verschelde, 1999), a software package which implements polyhedral homotopy continuation methods for solving systems of polynomial equations. The results of our simulations show that multiple solutions can occur. For example, for $N_1 = 11$, $N_2 = 5$, and the summary statistics

$$\begin{aligned} \bar{X} &= \begin{pmatrix} -1.5516 \\ -9.4713 \end{pmatrix}, & \tilde{S}_1 &= \begin{pmatrix} 0.3998 & -0.1026 \\ -0.1026 & 0.2378 \end{pmatrix}, \\ \bar{Y} &= \begin{pmatrix} -1.9175 \\ -10.4805 \end{pmatrix}, & \tilde{S}_2 &= \begin{pmatrix} 0.4193 & 0.0792 \\ 0.0792 & 0.0334 \end{pmatrix}, \end{aligned}$$

the real solutions for μ are

$$\begin{pmatrix} -1.3570 \\ -10.2957 \end{pmatrix}, \quad \begin{pmatrix} -1.2478 \\ -9.9902 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} -1.4451 \\ -9.6333 \end{pmatrix}.$$

This example seems, however, to be a rare exception. Indeed, we found that the bivariate Behrens-Fisher likelihood equations (3.1) had one real solution in about 99.5% of simulations, three real solutions in about 0.5% of simulations, and we found no instances in which the equations had five real solutions. However, it is possible that (3.1) has five real solutions when the data is generated from a "wild" distribution and not from the corresponding multivariate distributions. For instance, for $N_1 = 15$, $N_2 = 28$, and

$$\begin{aligned} \bar{X} &= \begin{pmatrix} -4 \\ -3 \end{pmatrix}, & \tilde{S}_1 &= \begin{pmatrix} 49.3619 & -45.0547 \\ -45.0547 & 42.4495 \end{pmatrix}, \\ \bar{Y} &= \begin{pmatrix} 4 \\ 1 \end{pmatrix}, & \tilde{S}_2 &= \begin{pmatrix} 52.8534 & 19.8380 \\ 19.8380 & 9.0472 \end{pmatrix}, \end{aligned}$$

the real solutions for μ are

$$\begin{pmatrix} 3.9822 \\ 1.0443 \end{pmatrix}, \begin{pmatrix} -3.7286 \\ 3.2906 \end{pmatrix}, \begin{pmatrix} -2.4192 \\ 4.6925 \end{pmatrix}, \begin{pmatrix} 2.0437 \\ 5.8993 \end{pmatrix}, \begin{pmatrix} 1.0089 \\ 8.2001 \end{pmatrix}.$$

To test for distinctions between the case of small and large samples in the bivariate case, we performed simulations in which N_1 and N_2 were randomly generated (uniform distribution) between 3 and 15. The outcomes are given as follows, with percentages rounded-off to two decimal places:

Please place Table 1 here

As noted above, none of these simulation resulted in five real solutions.

In the case of larger samples, our simulations resulted in the following outcomes:

Please place Table 2 here

Here again, no simulation resulted in five real solutions. (In both cases, the population mean μ is randomly generated from a uniform distribution on the subspace $[-20, 20] \times [-20, 20]$, and the population covariance matrices Σ_1 and Σ_2 are randomly generated in the manner described above, with positive diagonal entries whose values are no greater than 10.)

In summary, there seems to be little chance that a randomly generated, two-dimensional Behrens-Fisher problem will have three or more real solutions, and there is a high chance that it will have a unique real solution. The following supports the second conclusion for large sample sizes.

Theorem 4.1. *Suppose that the random samples X_1, \dots, X_{N_1} and Y_1, \dots, Y_{N_2} are drawn from independent normal populations $N_p(\mu, \Sigma_1)$ and $N_p(\mu, \Sigma_2)$, respectively. As $N_1, N_2 \rightarrow \infty$ the likelihood equations (2.4) for the Behrens-Fisher problem has a unique real root with probability one.*

Proof. If $\bar{X} = \bar{Y}$ then it follows from (3.2) that the *unique* real solution of the likelihood equations is $\hat{\mu} = \bar{X} = \bar{Y}$. Without loss of generality we can assume

that $\bar{X} = \bar{Y} = 0$, and with this the likelihood equations are

$$\frac{N_1 \tilde{S}_X^{-1} \hat{\mu}}{1 + \hat{\mu}' \tilde{S}_X^{-1} \hat{\mu}} + \frac{N_2 \tilde{S}_Y^{-1} \hat{\mu}}{1 + \hat{\mu}' \tilde{S}_Y^{-1} \hat{\mu}} = 0. \quad (4.1)$$

We argue that $\hat{\mu} = 0$ is a solution of multiplicity one for the system obtained by clearing denominators in (4.1). Let I be the ideal in $\mathbb{C}[\mu_1, \dots, \mu_p]$ generated by these p equations. The multiplicity of $\hat{\mu} = 0$ is the length of the artinian module

$$\frac{\mathbb{C}[\mu_1, \dots, \mu_p]_{\langle \mu_1, \dots, \mu_p \rangle}}{I \cdot \mathbb{C}[\mu_1, \dots, \mu_p]_{\langle \mu_1, \dots, \mu_p \rangle}}$$

over the local ring $\mathbb{C}[\mu_1, \dots, \mu_p]_{\langle \mu_1, \dots, \mu_p \rangle}$. I is generated by p polynomials given by $(N_1 \tilde{S}_X^{-1} + N_2 \tilde{S}_Y^{-1})\hat{\mu} + N_1(\hat{\mu}' \tilde{S}_Y^{-1} \hat{\mu})\tilde{S}_X^{-1} \hat{\mu} + N_2(\hat{\mu}' \tilde{S}_X^{-1} \hat{\mu})\tilde{S}_Y^{-1} \hat{\mu} = 0$. Each of these polynomials consists of a linear term and a cubic term. With probability one the rank of $N_1 \tilde{S}_X^{-1} + N_2 \tilde{S}_Y^{-1}$ over \mathbb{C} is p , and hence we can assume that I is generated by p polynomials of the form $\mu_i + g_i$ where g_i has degree three. This implies that the initial ideal of I in the local ring $\mathbb{C}[\mu_1, \dots, \mu_p]_{\langle \mu_1, \dots, \mu_p \rangle}$ with respect to the local term order *anti-graded revlex* as on page 152 of Cox, *et al.* (1998) is $\langle \mu_1, \dots, \mu_p \rangle$. By Corollary 4.5 of Cox, *et al.* (1998), we conclude that the length of the above module and hence the multiplicity of $\hat{\mu} = 0$ is one. Now as $N_1, N_2 \rightarrow \infty$, by the Law of Large Numbers, \bar{X} and \bar{Y} converge to μ , and S_1 and S_2 converge to S_X and S_Y . Since $\hat{\mu} = 0$ is the unique real solution to (4.1) with multiplicity one, and by the continuity of solutions to the general likelihood equations (2.4), we conclude that with probability one (2.4) has a unique real solution. \square

Acknowledgment. We thank Mathias Drton, Bernd Sturmfels, and the Editors for discussions and comments on initial drafts of this manuscript. Richards' work was supported in part by NSF grant DMS-0705210.

References

- Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, third edition. Wiley, New York.
- Behrens, W. U. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher*, **68**, 807–837.

- Buot, M.-L. G. and Richards, D. St. P. (2006). Counting and locating the solutions of polynomial systems of maximum likelihood equations, I. *J. Symbolic Comput.* **41**, 234–244.
- Catanese, F., Hosten, S., Khetan, A., and Sturmfels, B. (2006). The maximum likelihood degree. *Amer. J. Math.* **128**, 671–697.
- Cox, D. A., Little, J., and O’Shea, D. (1998). *Using Algebraic Geometry*. Springer, New York.
- Drton, M. (2007). Multiple solutions to the likelihood equations in the Behrens-Fisher problem. <http://arxiv.org/abs/0705.4516>.
- Drton, M. and Eichler, M. (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scandinavian Journal of Statistics* **33**, 247–257.
- Drton, M. and Richardson, T. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika* **91**, 383–392.
- Fisher, R. A. (1939). The comparison of samples with possibly unequal variances. *Ann. Eugen.*, **9**, 174–180.
- Hosten, S., Khetan, A., and Sturmfels, B. (2005). Solving the likelihood equations. *Found. Comput. Math.* **5**, 389–407.
- Kim, S.-H. and Cohen, A. (1998). On the Behrens-Fisher problem: a review. *J. Educational Behavioral Statist.*, **23**, 356–377.
- Linnik, Yu. V. (1967). On the elimination of nuisance parameters in statistical problems. *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability*, Vol. I: Statistics, pp. 267–280, Univ. California Press, Berkeley, CA.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, New York.
- Pachter, L. and Sturmfels, B. (2005). *Algebraic Statistics for Computational Biology*. Cambridge University Press, Cambridge, U.K.

- Stuart, A. and Ord, J. K. (1994). *Kendall's Advanced Theory of Statistics*, 6th edition. Edward Arnold, London.
- Sugiura, N. and Gupta, A. K. (1987). Maximum likelihood estimates for the Behrens-Fisher problem. *J. Japan Statist. Soc.*, **17**, 55–60.
- Szegő, G. (1939). *Orthogonal Polynomials*. American Mathematical Society, New York, NY.
- Verschelde, J. (1999). Algorithm 795: PHCpack: A general-purpose solver for polynomial systems by homotopy continuation. *ACM Trans. Math. Software*, **25**, 251–276.
- Wallace, D. L. (1980). The Behrens-Fisher and Fieller-Creasy problems. *R. A. Fisher: An Appreciation*, pp. 119–147. Lecture Notes in Statist., **1**, Springer, New York.

Table 1: Simulations with $3 \leq N_1, N_2 \leq 15$

Number of solutions	Frequency	Percentage
1	4450	99.29%
3	32	0.71%

Table 2: Simulations with $15 \leq N_1, N_2 \leq 60$

Number of solutions	Frequency	Percentage
1	4404	99.46%
3	24	0.54%

Department of Mathematics and Computer Science, Xavier University, Cincinnati, OH 45207.

E-mail: buotm@xavier.edu

Department of Mathematics, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132.

E-mail: serkan@math.sfsu.edu

Department of Statistics, Penn State University, University Park, PA 16802, and the Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709.

E-mail: richards@stat.psu.edu